Article Watch, April 2012

Clive Slaughter

Georgia Health Sciences University-University of Georgia Medical Partnership, Athens, GA

This column highlights recently published articles that are of interest to the readership of this publication. We encourage ABRF members to forward information on articles they feel are important and useful to Clive Slaughter, Georgia Health Sciences University-University of Georgia Medical Partnership, 279 William St., Athens GA 30607-1777. Phone: 706-369-5945; Fax: 706-369-5936; E-mail: cslaught@uga.edu; or to any member of the editorial board. Article summaries reflect the reviewer's opinions and not necessarily those of the association.

DNA SEQUENCING AND CHARACTERIZATION

Lam H Y K, Clark M J, Chen R, Chen R, Natsoulis G, O'Huallachain M, Dewey F E, Habegger L, Ashley E A, Gerstein M B, Butte A J, Ji H P, Snyder M. Performance comparison of whole-genome sequencing platforms. *Nature Biotechnology* 30;2012:78–82.

The two most commonly used platforms for complete genome sequencing, Illumina and Complete Genomics, use different sequencing technologies. The accuracy and completeness of variant calling by the two platforms are compared in this report. The genome of a single person is sequenced on each platform to an average depth of coverage of ~76×. Both platforms are shown to be capable of detecting most single nucleotide variants: 88.1% of the 3.7 million single nucleotide variants were concordant between the two platforms. However, platform-specific variants numbering in the tens of thousands were observed, including instances occurring in exons of 1676 genes. The detection of indels is subject to much larger platform-specific bias: only 26.5% of indels were concordant between the two platforms. Validation of single nucleotide variants by target enrichment authenticated 92.7% of the concordant single nucleotide variants and >60% of the platform-specific variants, indicating that they are genuinely present in the genome. These results indicate that comprehensive variant detection may best be approached by sequencing on both platforms if possible and validating by Sanger sequencing and array capture. Alternatively, the data may be supplemented with exome sequencing to make use of the greater depth of coverage to fill in gaps in coding regions.

doi:10.7171/jbt.2012-2301-005

Myllykangas S, Buenrostro J D, Natsoulis G, Bell J M, Ji H P. Efficient targeted resequencing of human germline and cancer genomes by oligonucleotide-selective sequencing. *Nature Biotechnology* 29;2011:1024–1027.

A variety of procedures has been devised for targeted resequencing, an application of genomic deep sequencing important in both clinical and basic science. This paper describes an implementation in which targets for resequencing are captured directly on an Illumina sequencer's flow cell, thereby eliminating intermediary steps requiring ancillary equipment. Target-specific oligonucleotides are first synthesized and immobilized on the flow cell to serve as both capture probes and sequencing primers. A single-adaptor library is then prepared from genomic DNA and added to the flow cell, where the desired targets are captured by the immobilized primer probes. Finally, the captured library fragments are prepared for bridge amplification, clustered, and sequenced. This sample preparation can be automated and can be completed in 1 day. The methodology is tested in a proof-of-principle study, in which exons of 344 cancer genes are resequenced from samples of genomic DNA from cancer or normal cells. The procedure is expected to be scalable to much larger numbers of genes.

Hindson B J, Ness K D, Masquelier D A, Belgrader P, Heredia N J, Makarewicz A J, Bright I J, Lucero M Y, Hiddessen A L, Legler T C, Kitano T K, Hodel M R, Petersen J F, Wyatt P W, Steenblock E R, Shah P H, Bousse L J, Troup C B, Mellen J C, Wittmann D K, Erndt N G, Cauley T H, Koehler R T, So A P, Dube S, Rose K A, Montesclaros L, Wang S, Stumbo D P, Hodges S P, Romine S, Milanovich F P, White H E, Regan J F, Karlin-Neumann G A, Hindson C M, Saxonov S, Colston B W. High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Analytical Chemistry* 83;2011:8604–8610.

Pinheiro L B, Coleman V A, Hindson C M, Herrmann J, Hindson B J, Bhat S, Emslie K R. Evaluation of a droplet digital polymerase chain reaction format for DNA copy number quantification. *Analytical Chemistry* 84;2012: 1003–1011.

Digital PCR is a methodology capable of providing absolute measures of nucleic acid concentration without the need for external calibration. A suitable dilution of the target DNA sample is partitioned into multiple replicate reactions, such that each reaction contains 0, 1, 2, 3, … molecules of the template. Amplification to the terminal plateau phase of PCR then permits discrimination of partitions containing one or more template copies from those containing none. Using Poisson statistics, the absolute number of target DNA molecules can then be calculated from the fraction of reactions yielding a positive response. Increasing the number of replicate reactions increases both the dynamic range and the precision of the measurement. The method is tolerant of wide variation in amplification efficiency, as its end-point is a binary positive or negative call for each partition. The two papers cited here describe a new addition to the available commercial implementations of digital PCR. This one uses water-in-oil droplets as the reaction partitions. A 20-μL mixture of sample and reagents is divided into 20,000 monodisperse droplets in which PCR amplification takes place. Workflows use conventional TaqMan and a 96-well plate format similar to real-time PCR. An automated droplet flow cytometer reads each set of droplets after amplification at a rate of 32 wells/h.

CARBOHYDRATES AND GLYCOCONJUGATES

Steentoft C, Vakhrushev S Y, Vester-Christensen M B, Schjoldager K T B G, Kong Y, Bennett E P, Mandel U, Wandall H, Levery S B, Clausen H. Mining the O-glycoproteome using zinc-finger nuclease-glycoengineered simple cell lines. *Nature Methods* 8;2011:977–982.

Mucin-type O-glycosylation, in which a glycan is linked to a serine or threonine residue via an *N*-acetylgalactosomine (GalNAc) moiety, is increasingly acknowledged to be an important regulator of protein function. Such O-glycan attachment is initiated by up to 20 GalNAc transferase genes in humans. This level of complexity has precluded the identification of consensus amino acid sequences for O-glycan attachment. To expand knowledge of O-glycan attachment sites, cell lines producing simplified, homogeneous O-glycan structures are developed here using zinc-finger nuclease-based targeting technology to ablate a protein essential for elongation of the O-glycosyl core. This results in truncated glycan moieties consisting only of GalNAcα-O-Ser/Thr or NeuAcα2-6GalNAcα-1-O-Ser/Thr O-glycans. Lectin weak affinity chromatography is then used to purify O-glycopeptides for amino acid sequencing by electron transfer dissociation mass spectrometry. More than 350 O-glycosylation sites on over 100 proteins are identified using this approach. The great majority of these sites was unidentified previously, including a GalNAc-O-glycan linkage to a tyrosine residue. The ability systematically to document O-glycan attachment sites is hoped to catalyze inquiry into the function of the complex repertoire of GalNAc transferases.

SMALL MOLECULE ANALYSIS AND METABOLOMICS

Heinzmann S S, Merrifield C A, Rezzi S, Kochhar S, Lindon J C, Holmes E, Nicholson J K. Stability and robustness of human metabolic phenotypes in response to sequential food challenges. *Journal of Proteome Research* 11; 2011:643–655.

Individuals may be distinguished on the basis of their urinary metabolic profiles. With the long-term goal of predicting the effect of dietary interventions to maintain or improve healthfulness of individuals, ^{1}H NMR spectroscopic profiling is used here to describe the effects of nutritional intervention on the urinary metabolic profile of seven individuals following a controlled, 7-day dietary protocol. The results confirm the existence of a personal, core metabolic phenotype that is stable to dietary modulation. However, day-to-day variability in some metabolites is strongly coupled to diet. These include microbial-mammalian co-metabolites, such as phenylacetylglutamine, 4-cresylsulfate, and indoxylsulfate, indicating that the microbiome responds rapidly to changing substrate patterns. Some metabolites are excreted universally in response to particular food challenges by all individuals, for example, tartrate, proline betaine, hippurate, and 4-hydroxyhippurate upon ingestion of fruit. These results suggest patterns of individual variability in response to diet, which will influence the pattern of future health-related studies.

MASS SPECTROMETRY

Lee J, Chen H, Liu T, Berkman C E, Reilly P T A. High resolution time-of-flight mass analysis of the entire range of intact singly-charged proteins. *Analytical Chemistry* 83;2011:9406–9412.

Arguing that the well-known peak-broadening, observed in time-of-flight mass spectrometers with proteins of increasing mass, results from expansion-induced kinetic energy dispersion, the authors describe a kinetic energy-reducing inlet system that produces high-resolution spectra for proteins across a broad mass range. Highly charged electrospray aerosol droplets are allowed to collide with the interior, grounded surface of a ¼-inch-diameter metal tube prior to passage through the source inlet orifice. This causes

protein ion charge reduction to yield singly charged ions and dissipates their expansion-induced forward momentum. The ions are captured in a large-radius linear ion trap and then passed into a smaller-radius linear ion trap, where they are collected in front of an end-cap electrode using a digital waveform technology. The ions are ejected on-demand into a time-of-flight mass analyzer as a tightly collimated ion beam with controlled kinetic energy. Proteins as large as myosin are detected with little diminution of resolution or sensitivity compared with protein of mass <20,000. The spectra show low noise levels, permit detection of noncovalent complexes and noncovalent solvent adducts, and support low nanosecond flight-time accuracy.

PROTEINS—PURIFICATION AND CHARACTERIZATION

Knispel R W, Kofler C, Boicu M, Baumeister W, Nickell S. Blotting protein complexes from native gels to electron microscopy grids. *Nature Methods* 9;2012:182–184.

Single-particle transmission electron microscopy of proteins requires very small amounts of sample and is permissive of some impurity. To avoid lengthy procedures for sample purification and preparation that yield more than the needed quantities for electron microscopy, a procedure is described, in which proteins are purified from mixtures by native PAGE and then blotted directly onto electron microscopy grids. Following electrophoresis, proteins of interest are located by staining a replicate lane with a protein stain and then aligning the elecropherogram with the unstained lane, allowing for gel shrinkage or swelling during staining. Glow-discharged copper grids, coated with a continuous carbon film, are then placed on the unstained gel at requisite positions, the interface between gel and grid is wetted with a droplet of electrophoresis buffer, and proteins are transferred by electroblotting. Enough protein typically remains in the gel for mass spectrometric identification. The upper size limit for protein complexes is dependent only on the pore size of the electrophoresis gel: particles as large as the 26 S proteasome (2.5 MDa) are recovered in sufficient yield for microscopy.

Zhou M, Morgner N, Barrera N P, Politis A, Isaacson S C, Matak-Vinković D, Murata T, Bernal R A, Stock D, Robinson C V. Mass spectrometry of intact V-type ATPases reveals bound lipids and the effects of nucleotide binding. *Science* 334;2011:380–385.

The authors successfully use electrospray ionization to translocate intact rotary ATPases from *Thermus thermophilus* (an archeon) and *Enterococcus hirae* (a eubacterium) into the gas phase for ion mobility and mass spectrometric analysis. The experimentally determined mass of the *Thermus* ATPase is 659,202 Da, which corresponds to the sum of the spectrometrically determined subunit masses, plus additional mass that results from incomplete desolvation and lipid and nucleotide binding. The data establish the stoichiometry of subunits in the complexes and permit ascertainment of the number and type of tightly bound lipid molecules in the complexes. The authors also show that gas-phase dissociation patterns are affected by nucleotide binding. These results reflect the regulatory action of nucleotides on ATP hydrolysis and proton translocation. This work contributes significantly to understanding of ATPase function and represents a milestone in the mass spectrometric study of protein complexes.

MICROARRAYS

Cretich M, Bagnati M, Damin F, Sola L, Chiari M. Overcoming mass transport limitations to achieve femtomolar detection limits on silicon protein microarrays. *Analytical Biochemistry* 418;2011:164–166.

Two experimental conditions for protein microarray experiments receive attention here to enhance fluorescence detection of antibodies in protein microarrays. First, glass slides are replaced by silicon with a silicon dioxide top layer functionalized by a co-polymer of *N,N*-dimethylacrylamide, *N*-acryloyloxysuccinimide, and 3-(trimethoxysilyl) propyl-methacrylate. This composition produces very low fluorescence background, and hence increases detection sensitivity. Second, incubation of the protein arrays with antibodies is performed with mixing in a horizontal shaker. Compared with static incubation, substantial increases in signal strength are observed.

PROTEOMICS

Ting L, Rad R, Gygi S P, Haas W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nature Methods* 8;2011:937–940.

Isobaric mass tags are molecules that can be coupled with peptides and used for relative quantification of peptides (and hence, their parent proteins) in samples mixed together for multiplexed mass spectrometric analysis. The tags fragment to produce reporter product ions of different mass, whose relative signal strength provides a measure of how much of each peptide came from each of the samples making up the multiplexed mixture. Although this approach to relative peptide quantification is convenient, it is also inaccurate, as many peptides cannot be isolated "cleanly" for mass spectrometric fragmentation. The presence of "contaminant" peptides that contribute to product ion spectra tends to diminish the true magnitude of expression differences between samples by making signal strength ratios regress to the median value of 1:1.

The extent of this problem is starkly demonstrated in the present paper, which demonstrates widespread ratio

distortion of this type. The authors label samples containing different quantities of the entire yeast proteome with six different Tandem Mass Tag reagents (Thermo-Scientific, Waltham, MA, USA) to produce three different reporter signal strengths after mixing the samples together. They then spike in three equal samples of the human proteome, which are labeled with three of the six mass tag reagents. Distortion of the expected signal strength ratios for the yeast peptides is attributable to the co-fragmenting human peptides. The authors find that the ratio distortion is not eliminated by excluding instances of peptides that appear likely to be mixed upon inspection of precursor ion spectra, or by improving chromatographic fractionation of the peptides before mass spectrometry, or by narrowing the isolation window used to trap precursor ions for fragmentation.

However, they do find that the problem is eliminated by the relatively simple expedient of quantifying reporter ions in MS3 scans. This is achieved by using peptides made by protein digestion with Lys-C, which allows the amine-reactive mass tags to attach to both ends of each peptide. In an LTQ Orbitrap Velos mass spectrometer, product ions derived from these peptides (which are labeled at one end only) are selected for higher energy collisional dissociation (HCD) to quantify the reporter ions. Using these two stages of fragmentation effectively eliminates the problem of acquiring product ion spectra from mixed precursors.

Wenger C D, Lee M V, Hebert A S, McAlister G C, Phanstiel D H, Westphall M S, Coon J J. Gas-phase purification enables accurate, multiplexed proteome quantification with isobaric tagging. *Nature Methods* 8;2011:933–935.

This paper addresses the same problem as Ting et al. above. Wenger et al. estimate that only 3% of peptide precursors are isolated for fragmentation without detectable contributions from other peptides, with consequent distortion of expression ratio measurements. However, they develop a different solution to the problem that relies on gas-phase purification of precursor ions. This is accomplished by peptide ion charge-reduction through gas-phase reaction with reagent anions. Arguing that precursors of similar mass-to-charge ratio (*m/z*) are unlikely to have the same *z* value, this charge-reduction reaction will yield ions of different *m/z*, which are readily isolated in pure form for HCD and quantification. This protocol increases quantitative accuracy substantially.

FUNCTIONAL GENOMICS AND PROTEOMICS

De S, Michor F. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nature Biotechnology* 29;2011:1103–1108.

Fudenberg G, Getz G, Meyerson M, Mirny L A. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nature Biotechnology* 29; 2011:1109–1113.

Genomes in many cancer cell types display complex alterations of copy number, including amplification, deletion, and rearrangement of genetic material. In cancers of 26 different types, more than 330,000 genome rearrangements have now been documented. The two reports cited here describe independent analyses of these somatic alterations, which point to an important role for the three-dimensional (3-D) arrangement of DNA in the nucleus in determining which alterations occur. The 3-D arrangement of DNA in the nucleus has previously been characterized by the Hi-C technology, in which neighboring strands of DNA are cross-linked, and the tethered strands are ligated together and subjected to paired-end sequencing across the ligated junctions to reveal the strands' original spatial proximity. Both reports indicate that Hi-C reads are enriched for sites subject to rearrangements in cancer. For example, the *BCR* and *ABL* genes are in nuclear proximity, although they reside on different chromosomes, and this proximity may be linked to the formation of the *BCR-ABL* fusion oncogene in leukemiagenesis. Furthermore, the 3-D architecture of DNA brings together genomic regions into "replication factories", within which DNA synthesis occurs simultaneously, and this organization contributes to determining the sequential pattern in which DNA from different genomic regions replicates during S phase. De and Michor note a correlation between cancer rearrangement sites and replication timing, which focuses attention on the process of replication in promoting chromosomal rearrangement. It is hoped that these insights, together with rapidly accumulating information about genomic rearrangements in cancer, will help shape detailed knowledge of the mechanism of genomic disruption in cancer.

Ingolia N T, Lareau L F, Weissman J S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147;2011: 789–802.

Ribosome profiling techniques are based on analysis of the fragments of RNA, ~30 bp in length, protected from RNAse digestion by the binding of ribosomes, which can be identified on a genome-wide basis by reverse transcription to DNA and deep-level sequencing. Significantly, these data define protein-encoding open reading frames (ORFs), which are notably difficult to catalog in complex genomes. With the use of the drug harringtonine, which pauses ribosomes at initiation start-sites, a wide range of unannotated or modified ORFs is documented in the present study of mouse embryonic stem cells. These ORFs

include the majority of a class of sequences previously designated "large intergenic noncoding RNAs" (lincRNAs), which now clearly contain highly translated, short ORFs, encoding "short, polycistronic, ribosome-associated RNAs" (sprcRNAs). The study further identifies over 1000 strong translational pause sites that may represent regulatory sites for translation. Moreover, alternate initiation sites are identified in many protein-coding genes. These are capable of producing extended or truncated forms of known proteins. Translation upstream of known ORFs (which occurs at sites that include non-AUG codons) could modulate the expression of the downstream-encoded proteins: upstream translation is shown here to decrease as the stem cells undergo differentiation. It is speculated that the density of ribosome footprints, observed by ribosome profiling, may provide a measure of the rate of protein synthesis, much as RNA-seq read-density measures mRNA abundance, providing a new dimension to analysis of the proteome.

CELL BIOLOGY AND TISSUE ENGINEERING

Dalerba P, Kalisky T, Sahoo D, Rajendran P S, Rothenberg M E, Leyrat A A, Sim S, Okamoto J, Johnston D M, Qian D, Zabala M, Bueno J, Neff N F, Wang J, Shelton A A, Visser B, Hisamori S, Shimono Y, Van De Wetering M, Clevers H, Clarke M F, Quake S R. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature Biotechnology* 29;2011:1120–1127.

Conventional studies of the progression of tumor phenotype, which rely on analysis of whole populations of cells, struggle to distinguish between changes that result from selection between genetically different cells arising from genome instability, and developmental changes that result from clonal heterogeneity due to epigenetic changes that may recapitulate differentiation processes in normal tissues. This paper characterizes the composition of colonic adenomas by study of gene transcription at the single-cell level. It uses PCR gene expression analysis of single cells sorted by flow cytometry. Gene expression changes characterizing cell subsets that arise in normal differentiation during epithelial cell turnover are first defined. Cell lineages arising in benign and malignant colorectal tumors are then characterized. The tumors are found to contain multiple cell types whose transcriptional patterns mirror those of the cell lineages in normal epithelium. Furthermore, tumor tissue generated from single cells can recapitulate the lineage diversity of the parent tumors. These results largely explain the heterogeneity of cells in these tumors as arising from multilineage differentiation rather than clonal genetic heterogeneity. Finally, tumors are classified according to their expression of genes characteristic of specific cell types, and this classification is shown to have potential prognostic value. The technology developed here is believed to have general value in profiling the cellular composition of solid tumors.

Fan Y, Rubakhin S S, Sweedler J V. Collection of peptides released from single neurons with particle-embedded monolithic capillaries followed by detection with matrix-assisted laser desorption/ionization mass spectrometry. *Analytical Chemistry* 83;2011:9557–9563.

New methods for collecting peptides released from single neurons in culture are developed in this study. Peptides are collected in capillaries of 250 μm i.d., positioned close to the target neurons. The capillaries contain a poly(steryl methacrylate-*co*-ethylene glycol dimethacrylate) monolith, which acts as a matrix into which are embedded extraction particles derivatized with pyrrolidone or ethylenediamine functional groups for peptide capture. Neurons from *Aplysia californica* maintained in artificial seawater are individually stimulated with 50 mM KCl. The collection capillaries efficiently bind peptides released into this high ionic-strength environment. The collected peptides are subsequently analyzed with MALDI mass spectrometry. This approach is anticipated to be useful for a broad range of neuropeptide experiments and can be adapted to collection of a variety of released molecules by appropriate choice of capture particles.

IMAGING

Kim J, Zhao T, Petralia R S, Yu Y, Peng H, Myers E, Magee J C. mGRASP enables mapping mammalian synaptic connectivity with light microscopy. *Nature Methods* 9;2012:96–102.

Genome reconstruction across synaptic partners (GRASP) is a fluorescence-based method for detecting proximity between neurons sufficiently close for functional synapses to form between them. The method works by functional complementation between two nonfluorescent split green fluorescent protein (GFP) fragments tethered to synaptic membranes of separate neuronal populations. The fragments reconstitute a fluorescent GFP at synaptic locations. GRASP was implemented previously for nematode and fruit-fly studies but is here re-engineered to match the 20-nm-wide synaptic clefts typical of mammalian synapses. Genes encoding GFP pre- and postsynaptic fragments are delivered to their target cell populations in mice by electroporation in utero or in adult mice by introducing viral vectors via stereotaxic surgery. It requires no genetic manipulation of animal models. Concordance of signals with the distribution of functional synapses is confirmed in well-studied brain regions and verified by electron microscopy. Absence of fluorescent signals at sites known to lack functional synapses is also confirmed, and expression of the GFP fragments is shown to produce no substantial change

in synaptogenesis. GRASP thus adds to the existing repertoire of methods available for mapping neuronal circuits in mammals.

BIOINFORMATICS

Müller T, Schrötter A, Loosse C, Helling S, Stephan C, Ahrens M, Uszkoreit J, Eisenacher M, Meyer H E, Marcus K. Sense and nonsense of pathway analysis software in proteomics. *Journal of Proteome Research* 10;2011:5398–5408.

The authors assess the usefulness of two commercially available pathway analysis tools for biological interpretation of proteomic datasets. The two packages are Ingenuity Pathway Analysis and STRING. Although the two tools use different algorithms, the information they provide is similar. To test the tools' ability to identify pathways, the authors input lists of proteins belonging to established pathways. Both tools are shown to identify apoptosis, tau phosphorylation, insulin signaling, amyloid precursor protein processing, and Wnt signaling pathways satisfactorily, when only proteins belonging to the requisite pathways are input. However, when irrelevant proteins are added to the input lists to simulate background proteins identified during proteomic analysis, the tools perform less well. Recommendations to users based on these results are formulated for investigators interested in using pathway analysis tools.